**INCIDENT IMPACT PREDICTION**

1. **Objective:**

The main objective of this project is to predict the impact and urgency of the customer problems with some of the pre-set attributes such as incident\_state, reopen\_count, sys\_mod\_count etc., Incident impact plays a major role in the services. Because of getting good customer servicing results. If 100 complaints are registered at a same time, It is not possible to solve all problems at a time. For this problems, impact and urgency plays a role. To improve the service’s and the time consumption of the services to make better, impact and urgency category plays a major role on that.

1. **Dataset Details:**

It contains 141712 rows and 36 columns. 36 columns shows that 36 attributes are in the dataset. They are ‘number’, ‘caller id’, ‘problem id’, ‘contact type’, ‘location’, ‘opened by’, ‘opened at’, ‘created by’, ‘created at’, ‘updated by’, ‘updated at’, ‘need sla’, ‘cmbd ci’, ‘knowledge’, ‘vender’, ‘notify’, ‘assigned group’, ‘assign to’, ‘category’, ‘subcategory’, ‘impact’, ‘urgency’, ‘priority’, ‘u priority confirmation’, ‘rfc’, ‘caused by’, ‘reassign count’, ‘reopen count’, ‘sys mod count’, ‘incident state’, ‘active’, ‘u symptom’, ‘resolved by’, ‘resolved at’, ‘closed code’, ‘closed at’.

In this attributes, ‘impact’ and ‘urgency’ are the dependent variable. Others are independent variable.

Attributes Details explanation:

1. Number: incident identifier (24918 different values)
2. Caller id: Identifier of the user affected.
3. Problem id: Identifier of the Problem associated with the incident.
4. Contact Type: Categorical attribute of the type of contact.
5. Location: Location of the Place affected.
6. Opened by: uses who reported the incident.
7. Opened at: reported date and time.
8. Created by: user who registered the incident.
9. Created at: registered date and time.
10. Updated by: user who update and generated the log record.
11. Updated at: updated date and time.
12. need sla: Boolean attribute that shows the SLA exceeded or not.
13. cmbd ci: identifier used to report the affected item.
14. knowledge: Boolean attribute shows whether knowledge document is used or not.
15. Vender: identifier of the vendor.
16. Notify: categorical attribute that shows whether notification was generated or not.
17. assigned group: Support group in-charged.
18. assign to: A person in in-charge.
19. Category: First level description of the incident.
20. Subcategory: Second level description of the incident.
21. Impact: Description of the impact (high, medium or low)
22. Urgency: Description of the urgency (high, medium or low)
23. Priority: calculated attribute that shows whether notification were generated or not.
24. u priority confirmation: whether the priority is double checked or not.
25. Rfc: (Request for change) identifier of the change request associated with the incident.
26. caused by: identifier of the RFC.
27. reassign count: how many times the incident has the group or the support analysis changed.
28. reopen count: how many times resolution was rejected by the caller.
29. sys mod count: number of incident updates until that moment.
30. incident state: 8 level of states that controlling the incident management.
31. Active: whether it is active or closed.
32. u symptom: User comments about service availability.
33. resolved by: identifier of the user who resolved.
34. resolved at: resolved date and time.
35. closed code: identifier of the resolution.
36. closed at: incident user close data & time.

From this, attributes except impact and urgency, we have to predict the perfect model for the impact and urgency attributes.

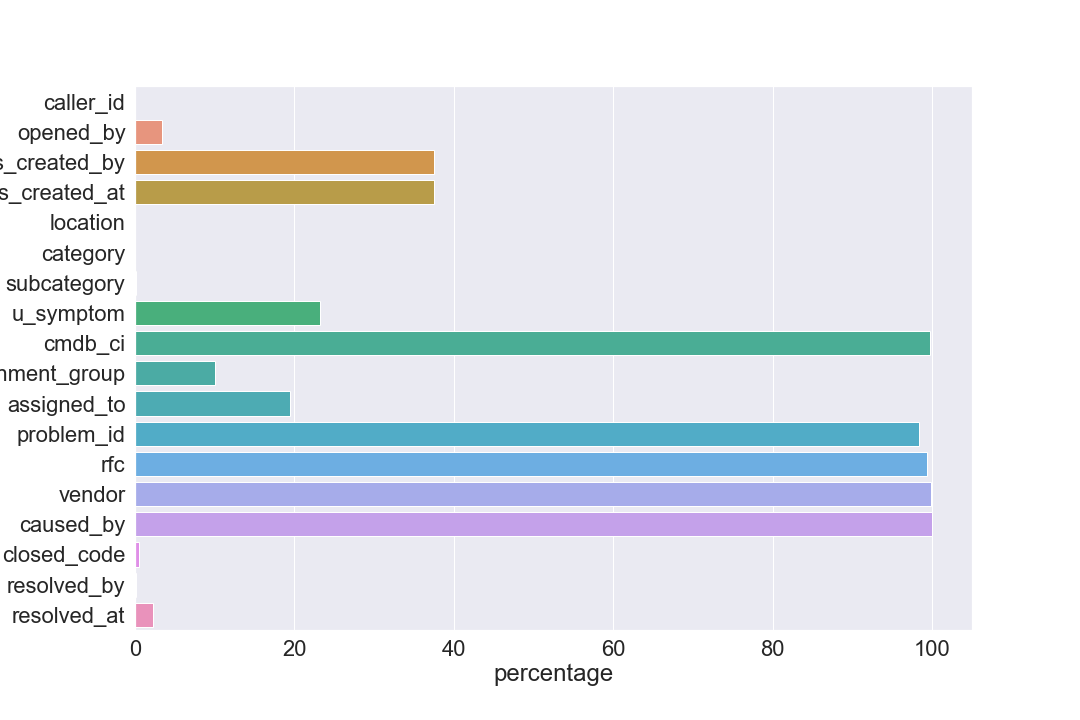
1. **Data Preprocessing:**

**Missing values:**

These data set contains no null values. Instead of that ‘?’ is present in the attributes.

There are 18 attributes with ?:

* ‘caller id’: 29(0.020% values are’?’)
* ‘location’: 76(0.053% values are ‘?)
* ‘category’: 78(0.055% values are ‘?’)
* ‘subcategory’: 111(0.078% values are ‘?’)
* ‘resolved by’: 226(0.16% values are ‘?’)
* ‘closed code’: 714(0.50% values are ‘?’)
* ‘resolved at’: 3141(2.22% values are ‘?’)
* ‘opened by’: 4835(3.41% values are ‘?’)
* ‘assignment group’: 14213(10.02% values are’?’)
* ‘assigned to’: 27496(19.40% values are ‘?’)
* ‘u symptom’: 32964(23.26% values are ‘?’)
* ‘sys created by’: 53076 (35.45% values are ‘?’)
* ‘sys created at’: 53076 (37.45% values are ‘?’)
* ‘problem id’: 139417 (98.38% values are ‘?’)
* ‘rfc’: 140721 (99.30% values are ‘?’)
* ‘cmdb ci’: 141267 (99.68% values are ‘?’)
* ‘vendor’: 141468 (99.82% values are ‘?’)
* ‘caused by’: 141689 (99.98% values are ‘?’)



**Why this ‘?’ is in the dataset:**

* **Caller id:**
  + It occur in only 3 number id’s(‘INC0016878’, ‘INC002463’, ‘INC0005690’) & 2 locations (188 & 71). Maybe, the problem occur in that place.
* **Opened by:**
  + Most of the cases are active cases(i.e.) out of 4800, 4120 is in active stage and also 1346 are new cases.
  + It exceeded the SLA, it may be a reason for “?”.
  + All are do not notify cases.
  + Impact and urgency are mostly medium. May be the problem is small.
  + Location most occur at 204, 161, 143, 108, 93, 51, 111, 55(Half of the unfilled rows occur in this location)
* **Sys created by and at:**
  + In this attribute also, Location most occur at 204, 161, 143, 108, 93, 51, 111, 179, 55 are unfilled. (Half of the unfilled rows occur in this location)
  + Impact and urgency are mostly medium. May be the problem is small.
  + Almost every case is exceeded SLA.
  + All attributes like ‘vendor’, ‘caused by’, ‘rfc’, ‘problem id’, ‘assigned to’, ‘cmbd ci’ are having ‘?’ and half of the ‘u symptom’ also ‘?’
* **Location:**
  + 90% of the ‘?’ in this location is new.
* **Category and Subcategory:**
  + Mostly New and even didn’t assigned to any group or servicers to handle the problem.
* **U Symptom:**
  + May be they don’t like to share the comments.
  + This also mostly occur in 204, 143, 161, 93, 108, 51, 179, 188, 55, 111 locations.
* **Assignment group:**
  + Occur in same 204, 143, 161, 93, 108, 51, 111, 179 locations.
  + Almost many cases are active than new, therefore something problem in that location handlers.
* **Assigned to:**
  + Same location handlers are problem. (204, 143, 161, 108, 93, 52, 179, 111)
* **Closed code:**
  + 607 is active, others are problem with the above specified location.
* **Resolved by and Resolved at:**
  + Some of the are active, others are mostly unfilled in red zoned locations.
* **Problem id:**
  + May or may not be generated. Because of caller id and location is enough.
* **Cmbd ci, RFC, vendor and caused by:**
  + Some simple reasons for not filling this rows are:
    - Not needed so much.
    - Cant able to remember the vendor.

From this, We must remove ‘problem id’, ‘rfc’, ‘vendor’ and ‘caused by’ column to improve our results. Other than that. ‘sys created at’, ‘sys created by’, ‘u symptom’ having greater than 20 percentage of ‘?’. Just try to replace that and perform the model. If it doesn’t work remove that also.

Others must replace with some other categories to maintain the data information.

For further process, we dropped the columns such as ‘problem id’, ‘rfc’, ‘vendor’, ‘caused by’ and ‘cmbd ci’. Because of having more than 99% of the data’s are ‘?’.

Other columns having ‘?’ are filled with the random values using the code.

‘’’’

def randomvar(df, var):

df[var + '\_fill'] = df[var]

random\_sample = data[data[var] != '?'][var].sample(len(data[data[var] == '?']))

random\_sample.index = data[data[var] == '?'][var].index

df.loc[data[var] == '?', var + "\_fill"] = random\_sample

for i in qstn\_name:

randomvar(data,i)

‘’’’

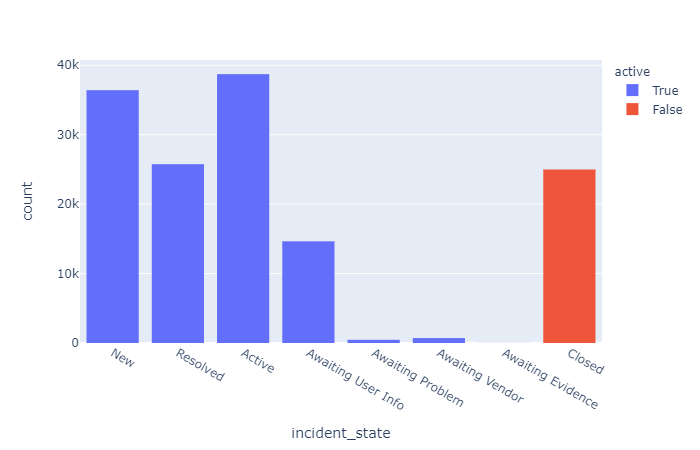
And then Preprocess the data by removing the extra words in the observations and also covert the string format of the datetime column convert to datetime format. These are the further preprocessing.

After the formation of pure data, we go to visualization steps.

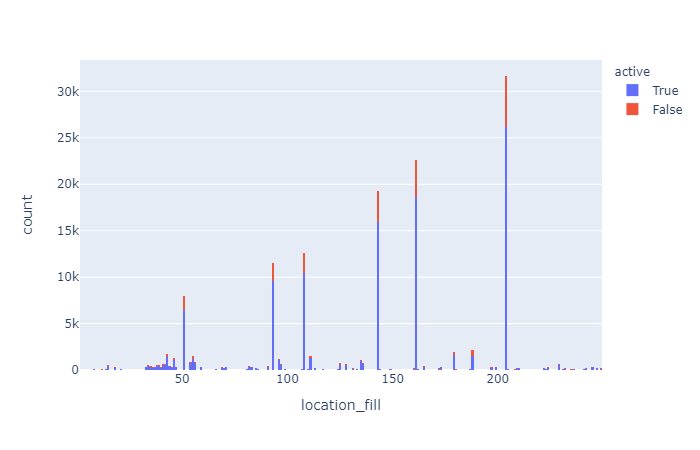
* **Incident state having wrong entity like ‘-100’:**
  + Replace that with some other state’s

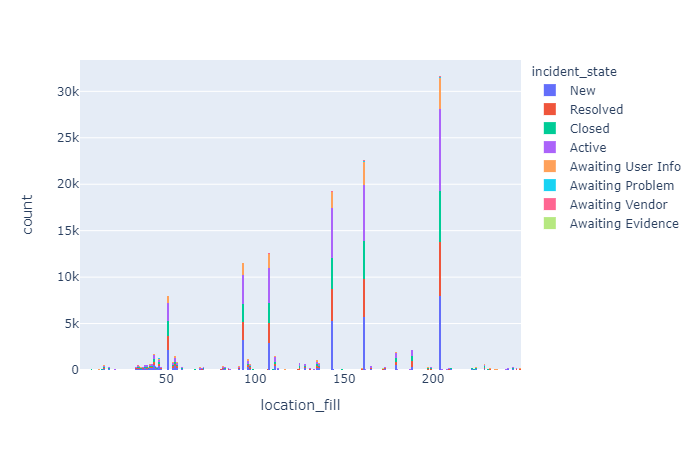
**Visualization:**

* By comparing “active” and “incident state”,

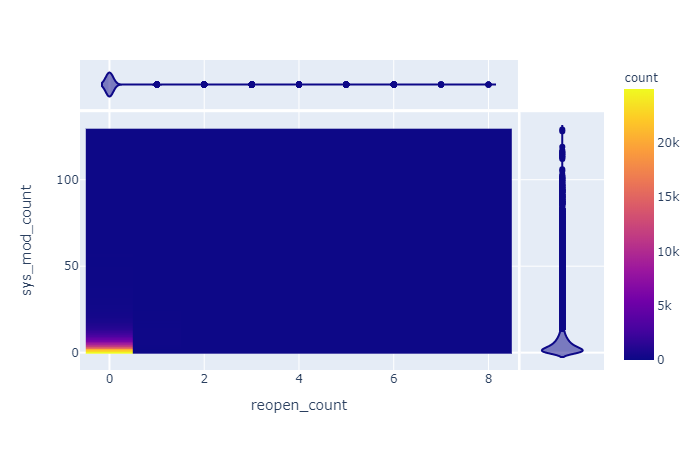
****

* + Most of them are in the active stages but in the various format’s like awaiting problem, awaiting evidence, awaiting vendor etc.,
* **Incident state & active based on location:**

****

****

* + problematic locations like 204, 143, 161, 108, 93, 52, 179, 111 having higher counts. It’s because of higher population in that area.
  + Everything looks good even at different stages.
* **Density heatmap between “reopen count” and “sys mod count”:**

****

* + It shows the higher number of updates leads to a smaller number of reopen count. That is, rejection of the resolution becomes less when the solution updates occur highly.